RARE JAROSITE DETECTION IN CRISM IMAGERY BY NON-PARAMETRIC BAYESIAN CLUSTERING

Murat Dundar *

Indiana University - Purdue University
Computer & Information Science Department
Indianapolis, Indiana, 46202 USA

Bethany L. Ehlmann

California Institute of Technology
Division of Geological & Planetary Sciences
Jet Propulsion Laboratory
Pasadena, California, 91125 USA

ABSTRACT

Discovery of rare phases on Mars is important as they serve as indicators of the geochemistry of the Mars surface and facilitate understanding of mineral assemblages within a geologic unit. Identification of rare minerals in high spatial and spectral resolution Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) visible/shortwave infrared (VSWIR) images has been a challenge due to the presence of both additive and multiplicative noise and other artifacts, affecting all collected images, in addition to the limited spatial extent of regions hosting these minerals. In an effort to automate this task we evaluate various clustering algorithms using the detection of rare jarosite, associated with spectrally similar minerals in CRISM imagery, as a case study. We compare non-parametric Bayesian and standard clustering algorithms and show that a recently developed doubly non-parametric Bayesian model could be effective for this task.

Index Terms— CRISM, jarosite, rare target detection, non-parametric bayesian, clustering

1. INTRODUCTION

Discovery of small, rare phases on Mars is important for two reasons. First, specific minerals such as alunite and jarosite (acidic), serpentine (alkaline, reducing), analcime (alkaline, saline), prehnite (200°C - 400°C), and perhaps phases yet to be discovered serve as direct environmental indicators of the geochemistry of waters on the Mars surface. Second, the identification of rare endmember phases facilitates understanding the mineral assemblages within a geologic unit, which are critical for identifying the thermodynamic conditions and fluid composition during interactions of rocks with liquid water.

The identification of these uncommon and spatially restricted mineral phases is difficult using existing CRISM processing techniques. The most common spectral mineral-identification method involves finding the ratio of the average spectra from two regions along-track in the image, where the numerator is the spectrum from the area of interest and the denominator is the spectrum derived from a spectrally homogeneous "neutral" region [1]. Spatial averaging filters additive random noise, whereas spectral ratioing reduces multiplicative noise. Summary parameters [2, 3] derived from key absorption bands are used to identify candidate regions for the numerator and denominator. Ratioing has proven very effective in the identification of minerals that occur over relatively large spatial extents in images. However, summary parameters have had limited success to date for identification of rare phases spanning a limited number of contiguous pixels in an image.

We consider rare jarosite detection as a case study and investigate the performances of several clustering algorithms toward automating rare phase detection in CRISM imagery. Jarosite is important because it is an indicator mineral for acidic, oxidizing conditions. Jarosite can be challenging to detect because its principal absorption occurs in a spectral region shared by absorptions in aluminum phyllosilicates and silica phases (Fig. 1) as well as iron-rich phyllosilicates, common mineral classes on Mars, which form spatially extensive geologic units [4].

2. METHODS

2.1. Image preprocessing and ground truth

Two images from Nili Fossae (FRT00009971, FRT0000A053) and one image from Mawrth Vallis (HRL000043EC) were used. Simple atmospheric and photometric corrections are applied to all three images using the CRISM Analysis Toolkit [7, 8]. Only the spectral channels that cover the spectral region from 1.0 to 2.6μm (248 channels) are used in this study. The channels corresponding to the remaining part of the spectrum (0.4-1.0μm and 2.6 to 4.0μm) were excluded because surface spectral properties at shorter wavelengths are

*This research was sponsored by the National Science Foundation (NSF) under Grant Number IIS-1252648 (CAREER). The content is solely the responsibility of the author and does not necessarily represent the official view of NSF

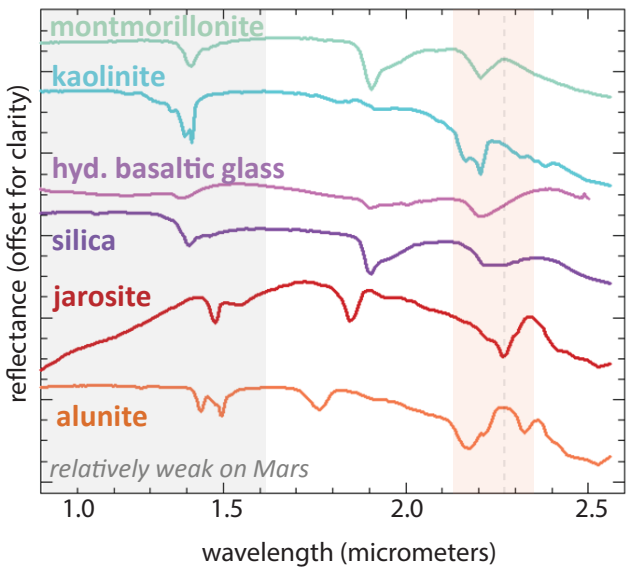

Fig. 1. Laboratory spectra of Mars-relevant minerals with absorptions from 2.17-2.27 μm. Jarosite is best distinguished from the other phases by its strong 2.2 μm feature with maximum absorption at 2.26-2.27 μm, distinctively longward of the others (red shaded region). Other absorptions are also distinctive and are sometimes detected in CRISM data, although vibrational absorptions due to H2O and OH are reduced in strength on Mars relative to their terrestrial laboratory measurements (gray shaded region). Spectra are from [5] and [6].

largely obscured by ubiquitous Martian dust and the longer wavelengths show low data quality and residual artifacts.

Jarosite locations are initially identified using the I^{2}GMM algorithm [9]. Because this algorithm is one of the algorithms used in this study for comparison, to avoid bias in establishing our ground truth, the final spatial extent of jarosite in each image is determined by visual inspection by investigating several ratioed spectra around this location and choosing the region that best recovers the distinctive doublet absorption in the 2.20-2.27μm region characteristic of jarosite in CRISM imagery. For all three images the extracted spectra are shown (Fig. 2).

Clustering experiments are performed with approximately 10K pixels obtained from 100 by 100 sub images including ground truth jarosite locations. Sub-images rather than original images are used for two reasons. First, for each algorithm several different parameter configurations are considered and for each configuration experiments are run ten times to measure variations in performance measure across different runs. Using a sub-image significantly reduces run-time for each algorithm. Second, for more accurate evaluation of the clustering algorithms, it is important to verify that all false positives identified by each algorithm are indeed false positives and not other jarosite regions that were overlooked when establishing the ground truth. This task can be more easily performed for a sub-image than the original larger image. Each pixel is projected onto the top ten principal components of the image I/F data. I/F data is derived from the radiance data by computing the ratio of the radiance to the solar irradiance at Mars [13]. Data points in the following presentation refers to pixels represented as points in this 10-dimensional vector space.

2.2. Standard clustering algorithms

The K-means algorithm and finite Gaussian mixture model (GMM) are popular algorithms for clustering data sets. K-means fixes the number of clusters K beforehand and assigns each pixel to one of the K clusters by minimizing a predefined distance metric in an iterative fashion. Similar to the K-means, finite GMM also fixes the number of clusters K beforehand, but unlike K-means fits a K-component Gaussian mixture model onto the data using the expectation-maximization algorithm [14]. K-means can be considered as a special case of GMM that restricts component covariances to spherical shapes.

2.3. Non-parametric Bayesian clustering algorithms

Finite GMM and K-means algorithms have two major limitations. First, no prior knowledge about cluster characteristics are used during clustering. Although lack of prior knowledge may not pose a serious problem for clustering data sets with balanced and well-defined clusters, prior knowledge may become critical when clustering data sets with rare clusters. Second, both finite GMM and K-means require that the number of components is defined in advance. Although there are several ways to predict the number of components in the data in an offline manner, these techniques are in general suboptimal as they decouple the two interdependent tasks: predicting the number of components and predicting model parameters.

A more flexible version of GMM can be derived by taking the limit over the number of mixture components to infinity. With this infinite version of GMM (IGMM) the actual number of components is automatically estimated during inference along with other component parameters in a single unified process [15]. IGMM is considered a non-parametric model as the number of components is no longer fixed during inference and can arbitrarily grow to better accommodate data sets as needed. It is also Bayesian as prior knowledge about number of clusters, cluster shapes and dispersion can be encoded into the model. [16].

$$\begin{aligned} \boldsymbol{x_i} &\sim P(\boldsymbol{x_i}|\theta_i) \\ \theta_i &\sim G \\ G &\sim DP(\alpha H) \\ H &= N(\mu|\boldsymbol{\mu_0}, \Sigma\kappa_0^{-1})W^{-1}(\Sigma|\Sigma_0, m) \end{aligned} \tag{1}$$

The IGMM model used in this study is shown in (1). According to this model individual data points $\boldsymbol{x_i} \in R^d$ are gen-

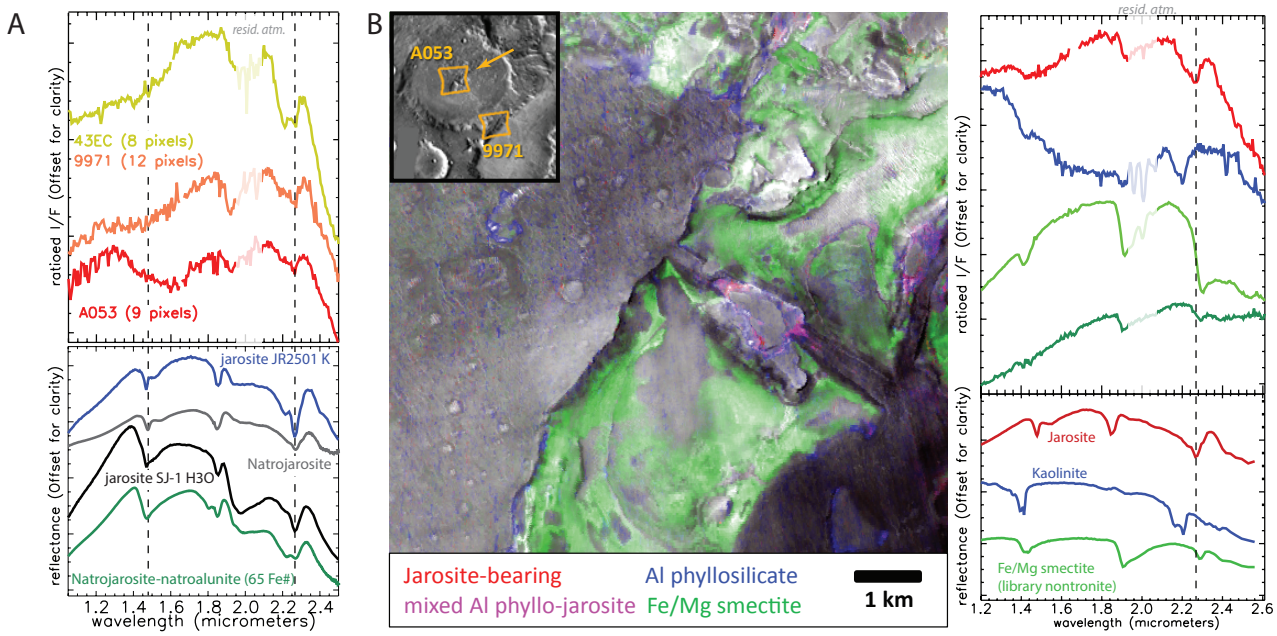

Fig. 2. (A) Averaged spectra for ground truth jarosite locations. Jarosite is identified primarily by the 2.26-2.27 μm absorption feature, though other absorptions are sometimes present. Library spectra from [5] and [10] are shown for comparison. The CRISM spectrum from image FRT000043EC is jarosite from an area in Mawrth Vallis previously reported in [11] to contain ferricopiapite, another indicator of acid sulfate alteration. FRT00009971 and FRT0000A053 are new detections from the Nili Fossae region. The context of FRT00009971 was previously described in [12]. (B) Follow-on ''expert user" analyses based on findings from FRT0000A053 confirm that jarosite-bearing terrains possess discrete spectral properties relative to other scene materials and permit spatial mapping of the phase in geologic context, here in association with an aluminum phyllosilicate-bearing stratigraphic unit. Colors correspond to the minerals indicated, with the figure generated by overlaying mapping of band strength at 2.24-2.27 μm (red); 2.28-2.31 μm (green) and 2.18-2.22 μm (blue) on a CRISM infrared albedo map. Library spectra from [5] are compared to spectra from regions in the CRISM image.

erated from Gaussian distributions whose parameters are in turn drawn from a Dirichlet Process (DP). A Dirichlet Process (DP) is a distribution over distributions. It generates random distribution G based on parameters H and α. The parameter H defines the base distribution from which the discrete probability masses, i.e., atoms, of G are drawn. The concentration parameter α changes the sparseness of G, which in turns affects the number of observed components in mixture models. The random measure G can be considered as a mixture of infinitely many atoms with their mixture weights drawn from a stick-breaking distribution

As data points are modeled by Gaussian clusters the base distribution H of DP serves as a prior over the cluster mean vectors $\boldsymbol{\mu_i}$ and covariance matrices Σ_i in which case θ_i denotes a set of two parameters, i.e., $\theta_i = \{\boldsymbol{\mu_i}, \Sigma_i\}$. In this case a conjugate distribution over a Gaussian data model would be a bivariate prior H that involves a Gaussian prior for the mean vectors and Inverse Wishart prior for the covariance matrices as in (1). We will denote this bivariate distribution by NIW.

NIW includes four hyperparameters: $\{\boldsymbol{\mu_0}, \kappa_0, \Sigma_0, m\}$. The hyperparameter $\boldsymbol{\mu_0}$ is the mean of the Gaussian prior defined over the cluster means. The hyperparameter κ_0 is a scaling constant that adjusts the dispersion of the cluster centers around $\boldsymbol{\mu_0}$. A smaller value for κ_0 suggests that cluster centers are expected to be farther apart from each other whereas a larger value suggests cluster centers closer to each other. The hyperparameters Σ_0 and m dictate the expected shape of the clusters. The minimum feasible value of m is equal to $d + 2$, and the larger the m, the less individual covariance matrices will deviate from the expected shape.

Both IGMM and GMM models each cluster with a single Gaussian component. This is a serious limitation if clusters emerge with multi-mode and/or skewed distributions, in which case IGMM creates additional components to better fit the data set. As there is no hierarchy in IGMM to allow for grouping of components into clusters, these additional components are treated as separate clusters, leading to suboptimal clustering performance. The infinite mixtures of infinite Gaussian mixture model (I^2GMM) is developed to circumvent this limitation by offering a two-layer non-parametric GMM [9]. This model is doubly non-parametric in terms of the number of clusters and the number of components for each

cluster, allowing for modeling clusters of various shapes. The generative model of I^{2}GMM is given in (2).

$$\begin{aligned} H &= NIW(\boldsymbol{\mu}_0, \Sigma_0, \kappa_0, m) \\ G &\sim DP(\gamma H) \\ \boldsymbol{\theta_j} = (\boldsymbol{\mu_j}, \Sigma_j) &\sim G \\ H_j &= N(\boldsymbol{\mu_j}, \Sigma_j/\kappa_1) \\ G_j &\sim DP(\alpha H_j) \\ \boldsymbol{\mu}_{ji} &\sim G_j \\ \boldsymbol{x}_{ji} &\sim N(\boldsymbol{\mu}_{ji}, \Sigma_j) \end{aligned} \tag{2}$$

According to this generative model a discrete mixing measure G is sampled from a global Dirichlet Process with base distribution H and concentration parameter γ. Cluster centers $\boldsymbol{\mu_j}$ and covariances Σ_j are drawn from G. Local DPs are defined, one for each cluster generated by the global DP. The base distributions H_j of local DPs are Gaussian distributions centered at $\boldsymbol{\mu_j}$ with a covariance of $\kappa_1^{-1}\Sigma_j$. Cluster-specific discrete mixing measures G_j are drawn from their corresponding local DPs with base distribution H_j and concentration parameter α. The mean vectors $\boldsymbol{\mu}_{ji}$ of components associated with cluster j are drawn from G_j. Data points are generated from Gaussian components with mean vector $\boldsymbol{\mu}_{ji}$ and covariance matrix Σ_j. In a nutshell this hierarchical model creates two level partitioning by clustering data points into components and components into clusters.

In addition to hyperparameters that already exist in the IGMM model, I^{2}GMM introduces two additional hyperparameters κ_1 and γ. The hyperparameter κ_1 models dispersion of component means around corresponding cluster center. A larger κ_1 leads to smaller variations in component means with respect to the corresponding cluster center, generating Gaussian-like clusters, whereas a smaller κ_1 dictates larger variations generating more flexible cluster shapes. The hyperparameter γ adjusts the number of clusters generated while α controls the number of components generated for each cluster.

3. RESULTS

We evaluate K-means, GMM, IGMM, and I^{2}GMM in terms of their detection accuracy. Each clustering algorithm returns cluster labesl for each pixel, which are used to compute F_1 score for each cluster. For each technique different clustering configurations are evaluated. K-means and GMM are run by varying K from 50 to 500 in increments of 50. For IGMM and I^{2}GMM we set α and γ equal to one, $\boldsymbol{\mu_0}$ to the center of the data, and $\Sigma_0 = (m-d-1)I/s$, where s is some scaling constant, I is the identity matrix, and d is the dimensionality of the data. The constant $(m-d-1)$ in Σ_0 is chosen to ensure that the expected Σ_k is equal to I/s. This leaves us with three free parameters (m, κ_0, s) to tune for IGMM for which twelve triplets are considered and four free parameters $(m, \kappa_0, \kappa_1, s)$ to tune for I^{2}GMM for which thirty six quartets are considered. These values are chosen to cover a broad range of cluster characteristics as discussed in [17]. For each clustering configuration F_1 scores are computed for all clusters and the cluster with the highest F_1 is considered as the most promising jarosite detection by each technique.

	K-means	GMM	IGMM	I^{2}GMM
9971	0.60/450	0.59/450	0.48/392	0.77/45
A053	0.46/300	0.44/500	0.46/195	0.70/34
43EC	0.86/400	0.63/400	0.78/190	0.83/101

Table 1. F1 scores and number of clusters K provided in the format F1/K. Results are averages over ten runs.

The results in Table 1 favors I^{2}GMM as a promising approach for rare jarosite detection in terms of both accuracy measured by F_1 score and efficiency measured by the number of clusters generated. Despite generating a significantly less number of clusters than other techniques, I^{2}GMM achieves the highest F_1 score for two of the images and produces an F_1 score that is comparable to K-means for the third image.

4. CONCLUSIONS

Automated rare phase detection in CRISM imagery is an important albeit a challenging problem that requires flexible models. A mixture model with a large number of components can more accurately estimate the density of the data. However, more accurate estimation of the density function may not improve clustering accuracy as the problem of many-to-one mappings between components and clusters remains unsolved in standard mixture models. I^{2}GMM addresses this limitation by jointly clustering data points into components and components into clusters in a unified model inference offering extreme flexibility in modeling a wide array of data distributions. Our preliminary results demonstrate the promising aspect of this framework for rare jarosite detection. Hyperparameter tuning is not specifically addressed in this study as we are still in the early stages of developing domain knowledge. An optimal parameter set that can more effectively encode domain knowledge can be obtained by processing several hundred images from different regions of Mars and performing both qualitative and quantitative evaluation of different hyperparameter sets. Additional details about technical aspects of I^{2}GMM and its implementation can be found in [17].

5. REFERENCES

[1] John F Mustard, SL Murchie, SM Pelkey, BL Ehlmann, RE Milliken, JA Grant, J-P Bibring, F Poulet, J Bishop, E Noe Dobrea, et al., "Hydrated silicate minerals on Mars observed by the Mars Reconnaissance Orbiter CRISM instrument," *Nature*, vol. 454, no. 7202, pp. 305–309, 2008.

[2] SM Pelkey, JF Mustard, S Murchie, RT Clancy, M Wolff, M Smith, R Milliken, J-P Bibring, A Gendrin, F Poulet, et al., "CRISM multispectral summary products: Parameterizing mineral diversity on Mars from reflectance," *Journal of Geophysical Research: Planets (1991–2012)*, vol. 112, no. E8, 2007.

[3] Christina E. Viviano-Beck, Frank P. Seelos, Scott L. Murchie, Eliezer G. Kahn, Kimberley D. Seelos, Howard W. Taylor, Kelly Taylor, Bethany L. Ehlmann, Sandra M. Wiseman, John F. Mustard, and M. Frank Morgan, "Revised CRISM spectral parameters and summary products based on the currently detected mineral diversity on Mars," *Journal of Geophysical Research: Planets*, vol. 119, no. 6, pp. 1403–1431, 2014, 2014JE004627.

[4] Bethany L Ehlmann, John F Mustard, Gregg A Swayze, Roger N Clark, Janice L Bishop, Francois Poulet, David J Des Marais, Leah H Roach, Ralph E Milliken, James J Wray, et al., "Identification of hydrated silicate minerals on Mars using MRO-CRISM: Geologic context near Nili Fossae and implications for aqueous alteration," *Journal of Geophysical Research: Planets (1991–2012)*, vol. 114, no. E2, 2009.

[5] R.N. Clark, G.A. Swayze, R. Wise, E. Livo, T. Hoefen, R. Kokaly, and S.J. Sutley, "USGS digital spectral library splib06a: U.S. Geological Survey, Digital Data Series 231," 2007, http://speclab.cr.usgs.gov/spectral.lib06.

[6] GA Swayze, RE Milliken, RN Clark, JL Bishop, BL Ehlmann, SM Pelkey, JF Mustard, SL Murchie, AJ Brown, Mro CRISM Team, et al., "Spectral evidence for hydrated volcanic and/or impact glass on Mars with MRO CRISM," *LPI Contributions*, vol. 1353, pp. 3384, 2007.

[7] Frank Morgan et al., "CRISM data users' workshop cat tutorial," `http://pds-geosciences.wustl.edu/missions/mro/CRISM_Workshop_090322_CAT_MFM.pdf`, March 2009.

[8] Scott L Murchie, Frank P Seelos, Christopher D Hash, David C Humm, Erick Malaret, J Andrew McGovern, Teck H Choo, Kimberly D Seelos, Debra L Buczkowski, M Frank Morgan, et al., "Compact Reconnaissance Imaging Spectrometer for Mars investigation and data set from the Mars Reconnaissance Orbiter's primary science phase," *Journal of Geophysical Research: Planets (1991–2012)*, vol. 114, no. E2, 2009.

[9] Halid Z Yerebakan, Bartek Rajwa, and Murat Dundar, "The infinite mixture of infinite gaussian mixtures," in *Advances in Neural Information Processing Systems*, 2014, pp. 28–36.

[10] Thomas M McCollom, Bethany L Ehlmann, Alian Wang, Brian M Hynek, and Thelma S Berquó, "Detection of iron substitution in natroalunite-natrojarosite solid solutions and potential implications for Mars," *American Mineralogist*, vol. 99, no. 5-6, pp. 948–964, 2014.

[11] William H Farrand, Timothy D Glotch, and Briony Horgan, "Detection of copiapite in the northern mawrth vallis region of mars: Evidence of acid sulfate alteration," *Icarus*, vol. 241, pp. 346–357, 2014.

[12] Bethany L. Ehlmann and Murat Dundar, "Are Noachian/Hesperian acidic waters key to generating Mars' regional-scale aluminum phyllosilicates? the importance of jarosite co-occurrences with alphyllosilicate units," in *Lunar and Planetary Science Conference*, 2015, vol. 46, p. 1635.

[13] S. Murchie et al., "Compact reconnaissance imaging spectrometer for mars (crism) on mars reconnaissance orbiter (mro)," *Journal of Geophysical Research: Planets*, vol. 112, no. E5, 2007.

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[15] Carl Edward Rasmussen, "The infinite gaussian mixture model," in *Advances in Neural Information Processing Systems*, 1999, vol. 12, pp. 554–560.

[16] Hemant Ishwaran and Lancelot F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. pp. 161–173, 2001.

[17] Halid Z. Yerebakan, Bartek Rajwa, Bethany L Ehlmann, and Murat Dundar, "The Infinite Mixture of Infinite Gaussian Mixtures for Clustering Data Sets with Multimode and Rare Clusters," `http://cs.iupui.edu/~dundar/papers/i2gmm.pdf`, November 2015.